ESSAY

SANDEL ON RAWLS

C. EDWIN BAKER†

In *Liberalism and the Limits of Justice,*[1] Michael Sandel offers an intriguing critique of John Rawls' *A Theory of Justice.*[2] Sandel's critique turns on his argument that "what issues at one end in a theory of justice must issue at the other in a theory of the person, or more precisely, a theory of the moral subject."[3] If from one direction the lens of the original position in *A Theory of Justice* shows us a moral theory, from the other direction it lets us see a "philosophical anthropology."[4] Sandel argues that Rawls' theory of justice requires that the person or moral subject be an abstract agent of choice, completely separate from her ends, personal attributes, community, or history. Only by adopting this notion of the person does Rawls' theory of justice make sense.

After describing the theory of the person to which he finds Rawls committed, Sandel claims that Rawls—and deontological liberalism[5] generally—fail because of the inadequacy and extreme individualism of this notion of the person. This individualism does not allow for the role of community in constituting the person, nor does it allow for the possibility that a person's meaningful identity is more a matter of cognition than choice. Sandel develops each objection into a major line of critique.

In the first critique, Sandel argues that the theory of the person to

---

† Professor of Law, University of Pennsylvania. B.A. 1969, Stanford University; J.D. 1972, Yale University. I have appreciated helpful and encouraging comments and criticisms on earlier drafts made by Drucilla Cornell, Robert Gordon, Peter Nordberg, and John Rawls. Although each will still disagree with some of what I say, I have benefitted from their comments.

[1] M. SANDEL, LIBERALISM AND THE LIMITS OF JUSTICE (1982).

[2] J. RAWLS, A THEORY OF JUSTICE (1971).

[3] M. SANDEL, *supra* note 1, at 47.

[4] *Id.* at 48.

[5] Sandel describes the core thesis of "deontological liberalism" as the view that the right is prior to the good, and that society "is best arranged when it is governed by principles that do not *themselves* presuppose any particular conception of the good . . . ." *Id.* at 1. In partial contrast, I have argued that a liberal legal order inevitably will favor some conceptions of the good over others and, furthermore, that this order will and should reflect both abstract principles of right and collectively favored conceptions of good. *See* Baker, *Outcome Equality or Equality of Respect: The Substantive Content of Equal Protection,* 131 U. PA. L. REV. 933 (1983).

which Rawls is committed is inconsistent with Rawls' difference principle.[6] The difference principle requires that basic societal institutions maximize the position of the worst off. Sandel claims that if the moral subject is an individual, then the difference principle will involve the conscription of some people's talents in order to benefit the worst off; the difference principle thereby treats those subjects as means. Only a group or community subject could both choose the difference principle and, since each person's talents would belong to this larger subject, avoid treating the moral subject as a means.[7] Thus, the Rawlsian theory of the moral subject as an individuated person is inadequate to support his theory of the right.

Sandel's second critique emphasizes that Rawls is committed to a thin, denuded notion of the person—a person separate from all ends, commitments, and capacities. This self is so sparse that it cannot constitute an object for self-reflection. It can only be a subject that is, at most, capable of arbitrary and ultimately meaningless choice. The arbitrariness and meaninglessness of this choice result in another fault—an inadequate theory of the good. In combination these two critiques argue that Rawls' notion of the person is neither appealing, consistent with our understanding and experience of ourselves, nor adequate to support Rawls' theory of justice. Specifically, the Rawlsian theory is inconsistent with selves who are constituted by their values, character, commitments, and practices, who are partially constituted by their membership and participation in communities, or who engage in deep self-reflection.

In Part I, I explain why I believe Sandel's description of a Rawlsian anthropology is wrong. Rawls undertakes only to derive the limits that justice would impose on acceptable frameworks for human interaction. To do so, he need only postulate certain universal qualities that we do or should attribute to the person, or to acceptable human interaction. Rawls only needs a theory of those aspects of the person or of human interaction that are relevant to his enterprise. Of course, he presumably should defend his implicit claim that those aspects rather than other or all aspects are relevant. Sandel's error, however, lies in assuming that those few universal qualities that Rawls emphasizes reflect a complete Rawlsian theory of the person.

Although not explicitly developed in his book, Sandel constantly promotes the notion of a group subject. In turn, this suggests the idea of group or collective rights that cannot be derived from, or limited by, individual rights. In this regard, I begin in part I the argument that

---

[6] *See* M. SANDEL, *supra* note 1, at 70-72, 76-78, 96-103, 139-41, 148-52, 178.
[7] *See id.* at 70-72, 76-78, 80.

Rawls' emphasis on individual rights is more appealing.

Parts II, III, and IV consider Sandel's more specific challenges to Rawls' theory. Part II critiques Sandel's claim that the difference principle requires a group subject; Part III critiques Sandel's claim that Rawls' apparent acceptance of a moral or preinstitutional basis for retributive, but not for distributive, justice shows that Rawls is confused or inconsistent; Part IV challenges the claim that justice may in certain circumstances be a vice, rather than having the primacy that Rawls asserts. In each part, I will show that Sandel's challenge to Rawls fails, and that Sandel's own position is either unnecessary or has ethically unappealing implications. The discussion in each section points, however, to a more fundamental issue that Sandel never explicitly addresses but on which the disagreement between Sandel and Rawls may depend: the acceptability of what I call a "two-level political theory."

The two-level theory assumes that universal attributes exist, either for human beings or for human interaction, and are distinguishable from a second level of attributes peculiar to each of us. The theory then claims that these universal attributes carry meaningful implications for the determination of a just social order and that these implications have a constitutive priority over the implications that can be derived from second level attributes. In Part V, I suggest that my present defense of Rawls would collapse, and aspects of Sandel's argument would succeed, if reliance on such a two-level theory were unacceptable. A full exploration and defense of the two level theory is beyond the scope of this essay. Nevertheless, I offer several reasons to think that reliance on it is acceptable, thereby suggesting the propriety of Rawls' general approach.[8]

I think that much of Sandel's image of the person and emphasis on community is right and important. Nevertheless, I believe that his argument ultimately fails as a critique of Rawls, and that Sandel's account of the group subject could provide the basis for a dangerous and unwarranted notion of group or community rights.

I. The Denuded or Abstract Person

A. *Sandel's Anthropology*

According to Sandel, Rawls' theory of the person depicts an agent of choice that is separate from her ends, attributes, talents, values, char-

[8] This approach is also implied in my earlier articles on constitutional theory. *See, e.g.*, Baker, *Scope of the First Amendment Freedom of Speech,* 25 UCLA L. Rev. 964 (1978); Baker, *supra* note 5.

acter, and commitments—that is, separate from all the particular or individual qualities that she may *possess*. The agent may possess these qualities but they do not even partially *constitute* the person.[9] The traits are "mine rather than me."

Two important corollaries follow from this conception of the person. First, since this prior-individuated agent is barren of all concrete features, she cannot, in any way, be constituted by community. As Sandel describes Rawls' theory: the plurality of individuals is prior to their unity.[10] Second, this Rawlsian person is incapable of self-reflection or of cognition in the sense of learning or comprehending who she is. This incapacity directly reflects the nature of the Rawlsian person. Since learning, comprehending, or discovering one's identity implies a basis of identity independent of choice, any role for this self-reflective or cognitive capacity would be inconsistent with the Rawlsian person whose defining feature is her freedom or capacity to choose her identity. Also, since the essential agent is completely denuded of all personal attributes, there is no substance left about which she could self-reflect. Both this inherent separateness of the person from community, and the fundamental defining quality of individual autonomous choice, make Rawls' notion of the person inconsistent with any constitutive role for community.

Sandel is persuasive in arguing that the thin, denuded image of the person or moral subject that he attributes to Rawls is inconsistent with our experience and understanding of ourselves, and with our aspirations. The question remains, however, whether Rawls is committed to this image of the person as a general account of what we are. I will argue that Rawls constructs this limited conception of the person for a limited but appropriate purpose and that this enterprise does not commit him to any general account or theory of the person.

Questions about the proper constitution of a framework of interaction necessarily present different issues than do questions about proper conduct within that framework. In some respects these differences are obvious. Reference to some sort of framework is needed to identify and justify a full set of rights (that is, claims addressed to others), or expectations, legitimate or otherwise, that create and secure, but also bound, a person's opportunities for action and choice. Questions concerning the permissibility of conduct within this constituted framework can rely upon the rights or expectations created by the framework in a way that questions concerning what framework we ought to have cannot. Within

---

[9] *See* M. SANDEL, *supra* note 1, at 20, 55-58.
[10] *See id.* at 50-51.

history and within a community, we expect that a person's actions, choices, and self-reflection normally will reflect her nature as a full person, "thick" with values, attributes, and qualities. The problem is that this person is not relevant for Rawls' critical project. Claims justified solely within and by reference to a framework of interaction can neither justify nor critique the framework itself. Likewise, the features of persons who make these claims and, more importantly, who are partially formed by that framework do not provide an adequate standpoint from which to justify or critique the framework. Or, at least implicit in Rawls' enterprise is the claim that the "thick" person, fully constituted by all her history and talents, does not provide an adequate standpoint from which to evaluate the framework. Rawls, therefore, must construct an alternative standpoint—the standpoint of the abstract moral agent of the original position. Thus, it is not a flaw but an objective of his approach that we do not identify with the conception of the person constructed for purposes of deriving critical principles.

For Rawls, and possibly for deontological liberalism generally, a very limited but very general or universal set of aspects that we attribute to the person or to human interaction has particular relevance to the search for basic principles of social organization. These general aspects, however, need not be the ones that are most important to what a fully constituted person takes herself to be. The claim that these aspects lead to basic principles of social organization only means "basic" in a trumping or priority sense, not that these principles are necessarily experientially most important or most burdensome. Rawls need only claim that some aspects of our notion of the self or of human interaction are relevant, and others are not relevant, to some basic issues concerning the proper framework of interaction.[11]

---

[11] Often, my discussion of Rawls will emphasize how I understand and would accept, to the extent that I do, his main moves in A THEORY OF JUSTICE, *supra* note 2, rather than describe how I think Rawls understands his own moves. Moreover, I will not claim that my remarks will be consistent with all of Rawls' more recent commentary. Aspects of my interpretation, however, are supported by a careful reading of some of his remarks at the Dewey Lectures, reprinted in Rawls, *Kantian Construction in Moral Theory*, 77 J. PHIL. 515 (1980). Consistent with Sandel, Rawls emphasizes throughout these lectures that his theory of justice relies upon a particular conception of the person. Rawls makes clear, however, that this conception of the person is not intended to be a complete conception, but is a conception that is specifically "connected with society's public conception of justice." *Id.* at 545. He notes that "within different contexts," (I would say, for different purposes), "we assume diverse points of view toward our person." *Id.* Rawls later emphasizes how, for different issues, we will have and use distinguishable concepts of the self. *Id.* at 571. He only needs a conception appropriate for a theory of justice. Thus, in evaluating Rawls' argument, the issue should be whether his concept of the person is adequate for this purpose, not whether it is adequate for our self-understanding in some other or broader context.

The Rawlsian claim that from an assumption of general qualities of humans or of human interaction we can arrive at conclusions concerning the proper framework of interaction, should be controversial and needs exploration. But Sandel's notion that a backwards look through the original position will show a Rawlsian theory of the person seems quite strange. A major purpose of the original position is to emphasize the view that many aspects of who we are ought to be irrelevant to certain issues—and specifically, ought to be irrelevant to the derivation of Rawls' principles of justice.

If Rawls' claim is right, then in arguing for a theory of justice, he can disregard many aspects of human interaction or of the nature of the person. He can properly deny relying on any general and complete theory of the person. The features that Sandel derives by looking in reverse through the original position could, at best, consist only of those specific, limited aspects of the moral agent, or of human interaction, that Rawls claims are relevant for the limited enterprise of deriving general principles of justice. Moreover, persuasively derived principles of justice may not determine, but may merely constrain the constitution of acceptable frameworks of interaction by ruling out certain unjust elements.[12] If so, then Sandel's philosophical anthropology will find only those aspects of the person that justify limits on the choice of frameworks.

It should be noted that, although not crucial to my present argument, even the notion that Sandel would correctly discover some general aspects of the person might be a mistake. I generally equivocate in this essay between two interpretations of the theoretical basis of Rawls' approach. Even if most students of Rawls see him as clearly relying on a Kantian theory of the person, or a philosophy of consciousness, I find a second interpretation, which views both the individual and individual consciousness to be an aspect or product of interaction, to fit his argument as well or better than the first interpretation.

The general features that justify Rawls' principles of justice can constitute either a partial theory of the person or a partial theory of human interaction. Since Rawls' enterprise is to find the principles that constrain proper frameworks for group interaction, one might expect that aspects of *interaction,* not aspects of an isolated agent, would be the proper focus. In that case, Rawls need posit nothing about the existence or nature of the person in isolation. Group interaction would be

---

[12] *See* Baker, *supra* note 5, at 943-72. In my earlier paper, I criticized Rawls' derivation of the difference principle and, using his methodology, argued for alternative principles. That paper, however, relied on those aspects of Rawls' approach that are defended here against Sandel's more fundamental critique.

key. Thus, contrary to Sandel's charge that Rawls places plurality prior to unity,[13] Rawls need investigate only the proper respect for plurality (that is, for individuals) that should be given within unity. In this interpretation, Sandel's look backwards would find only aspects of a theory of human interaction and, possibly, the way that interaction assumes a role for an autonomous moral agent.

This interpretation of Rawls' enterprise makes it directly parallel, at a more overtly political level, to Jürgen Habermas' effort to find conditions of interaction in which attempts to reach agreement through noncoerced communicative action can succeed.[14] Both enterprises assume that justifiable interaction requires that the circumstances of interaction meet certain conditions. Rawls views his task as finding principles that represent people as equals—clearly an emphasis on relationships between people—and attempts to find criteria for evaluating a framework of interaction that embodies this claim. Yet Rawls also emphasizes people as rational and autonomous agents concerned with advancing their individual interests. Despite this second emphasis, Rawls does not imply or assume that this conception of the person is empirically or historically accurate, or even that it is a relevant conception of the person for other purposes. Thus, his approach is less a hypothesis about people's "nature" than a hypothesis about how the social order ought to view them. This conception of the person for the purpose of a theory of constitutive social principles is an assumption about how we must view the other in order to engage in moral activity or, possibly, in undistorted communicative interaction. Thus, particular aspects of human interaction, not of the isolated person, may provide the motor for Rawls' approach. If so, then Rawls' enterprise is more related to a philosophy of communicative interaction and intersubjectivity than to a philosophy of consciousness.[15]

Under this second interpretation of Rawls' enterprise, we must recharacterize what is seen when looking backwards through the original position. We would not see a theory of the person at all, either in whole or in part. Rather, we would see a theory about the few premises of human interaction that must be realized if we are to engage in equal

---

[13] *See* M. SANDEL, *supra* note 1, at 50-51.

[14] *See* 1 J. HABERMAS, THE THEORY OF COMMUNICATIVE ACTION: REASON AND THE RATIONALIZATION OF SOCIETY (1981) [hereinafter cited as J. HABERMAS, COMMUNICATIVE ACTION]; J. HABERMAS, COMMUNICATION AND THE EVOLUTION OF SOCIETY (1979).

[15] *See* J. HABERMAS, COMMUNICATIVE ACTION *supra* note 14. *See also* R. BERNSTEIN, BEYOND OBJECTIVISM AND RELATIVISM: SCIENCE, HERMENEUTICS, AND PRAXIS (1983); R. RORTY, PHILOSOPHY AND THE MIRROR OF NATURE (1979). *Cf.* M. SANDEL, *supra* note 1, at 53 (criticizing Rawls for placing plurality prior to unity).

and uncoerced interaction. Specifically, this theory would assert that justifiable structures of interaction must respect people as equals and as autonomous or choosing agents. The claim is not that this is how people always are, or even that we ought to view people this way for all purposes. The claim is only that for purposes of morally acceptable interaction, in the framework that creates specific expectations or rights, people ought to be respected as this type of being. Nevertheless, regardless of whether one accepts this second philosophy of communicative interaction, or the first philosophy of consciousness interpretation, my key point is that looking backwards through the original position will only reveal a *partial* theory—whether of the person, or of egalitarian human interaction.

Once the original position methodology is seen as not implying a complete theory of the person, flaws in many of the more specific interpretations that Sandel attributes to Rawls' theory of the person become evident. For example, Sandel emphasizes Rawls' statement that " '*the self is prior to the ends which are affirmed by it.*' "[16] Sandel argues that this assumption about the priority of the self over its ends commits Rawls to a voluntarist, rather than a cognitive, notion of agency.[17]

Rawls' statement, however, need only imply that our notion of the self is not reducible to its ends, or more specifically, that the self may properly claim and exercise responsibility for its ends. In this sense of being nonreducible, the self is prior to its ends. Rawls need not exclude the possibility that, in another sense, basic ends partially define the fully constituted or "thick" self. The idea of freedom, expressed in the notion of choosing who and what we will be, implies both that ends are central to our being and that an aspect of our being is involved in choosing or affirming those ends.

Likewise, Rawls' notion of the self as an agent capable of choice does not rule out fundamental self-knowledge through cognition and self-reflection. To explain the role of principles of justice that respect individual autonomy, Rawls need only argue that people are involved in choosing or affirming some important aspects of the self and that basic social institutions ought to respect this role of agent responsibility. Moreover, Rawls can make a broader argument. As long as those aspects of the self that are discovered through self-reflection are not inconsistent with the idea of choosing them had they been subject to choice (that is, if the discovered aspects do not require behavior or institutions prohibited by the principles of justice), the principles of justice

---

[16] M. SANDEL, *supra* note 1, at 19 (quoting J. RAWLS, *supra* note 2, at 560) (emphasis added by Sandel).

[17] *See* M. SANDEL, *supra* note 1, at 58-59; *see also id.* at 22, 53-57, 133, 179.

would be consistent with cognition as well as choice.[18] For example, a social structure that respects people as agents who can rightfully choose to be Muslim is consistent with people discovering through cognition that their essential nature is defined by being Muslim. The only way in which Rawls' emphasis on agent choice or responsibility could be inconsistent with a cognitive notion of agency is if cognition were to reveal that the person's essential nature is to be forced or manipulated into being something, for instance, a Muslim, that she does not wish to be.

The cognitive notion of agency even corresponds to Rawls' scheme in an important respect. The original position itself is a device to aid self-reflection.[19] Rawls assumes that self-reflection helps achieve intelligent understanding and commitment to the principles of justice. His notion of the subject views the person as rightfully exercising responsibility for accepting or denying what she learns through self-reflection. (In this sense, Rawls could argue that his methodology corresponds to a Kantian notion of freedom.) The Rawlsian approach rejects only the conclusion that the cognition of someone other than the self can be an appropriate basis to define or limit the self—although the other's cognition or choice often does have an appropriate role in political or collective choice.

Thus, Rawls can remain agnostic both about whether our ends originate in choice or are discovered through self-reflection, and about the degree to which we are constituted by our ends and our community. He uses a sparse, voluntarist conception of the self, but only for purposes of thinking about the propriety of constraints on the structure of interaction, that is, for purposes of thinking about principles of justice. This use reflects Rawls' implicit claim that the propriety of people's interaction depends on treating the self both as the key unit of concern and as an agent that can properly assert a concern with, and claim some responsibility for, who she is (even if that involves her acceptance

---

[18] *See* Rawls, *supra* note 11, at 543-44 ("Claims that are said to be founded on duties to self, if some hold that there are such duties, are counted as self-originating for the purposes of a conception of social justice.").

[19] Arguably, the original position can best be seen as a rhetorical device designed to help us develop or understand, not choose, principles of justice. As Sandel belatedly notes, the original position is not a place for agreement between people or a place for unreflective or noncognitive choice. *See* M.SANDEL *supra* note 1 at 128-29. Rather, the original position is a device to further the cognitive activity of reflection. The notion of *plurality* implicit in the idea of contract does not relate so much to a proposed activity of parties agreeing. Rather, "contract" emphasizes that the object of cognition is the proper response to and constitution of our plurality. The contract image also embodies the assumption that the proper constitution must respect our autonomy; finally, a contract's quality of being binding embodies the notion that we ought to be committed to supporting in our practice the results of our reflections.

of an identity that she discovers in history and community). Thus, the structure of interaction must embody the assumption that people should be treated as having in common a capacity for choice and for responsibility—and the structure must be consistent with respect for these abstract aspects of the person. Only such a structure exhibits concern for and gives scope to self-reflection or to noncoercive dialogue as a means for either self-knowledge or self-choice.

Sandel's recognition that a person can have meaningful existence only within a historically grounded community, and that the community will constitute many particular aspects of that person's identity, is not inconsistent with Rawls' reliance on a sparse, featureless agent of choice for purposes of deriving the principles of justice. The only conception of the person that would be inconsistent with Rawls' approach would be a conception that would justify coercively subordinating the individual for the sake of realizing the person or advancing some larger being, agent, or idea. Typically, this conception would treat the self as being subordinate to a group, which would then have rights over the self. Only in the sense of rejecting this view of the self as properly subordinate to the group is Rawls fundamentally individualistic. Rawls could accept a view that a person is basically constituted by her membership in a group, but he probably would assert that the group is valuable and significant because it is important to the individual. The individual is the entity that can make ethical claims. Only this view, even as it admits that the individual cannot be conceived of as separate from the group, treats the individual's acceptance—her choice—of history, tradition, or the group as important. In contrast, the view suggested by Sandel's notion of a group subject, is entirely consistent with the coercive imposition of the particularities of the group on the person and with the coercive subordination of the individual.[20]

Sandel quotes Rawls' statement that "*[t]he essential unity of the self is . . . provided by the conception of right.*"[21] Sandel goes wrong, however, by interpreting "essential unity" to mean "entirety" or "most important aspect" of the self. Rawls only emphasizes those aspects of the self that are relevant to his objective of finding appropriate principles of justice. Rawls does not attempt to derive a full theory of society—which might require a full theory of the self. Properly developed, his approach does not lead to principles of justice that completely determine the framework of a just society. Rather, they merely distinguish

---

[20] The fear that Sandel's analysis leads in this direction clearly animated Brian Barry's response. *See* Barry, Book Review, 94 ETHICS 523, 525 (1984).

[21] M. SANDEL, *supra* note 1, at 21 (quoting J. RAWLS, *supra* note 2, at 563)(emphasis added by Sandel).

just from unjust societies.[22] Likewise, his principles do not determine, but only constrain, how people should develop themselves within a just framework. The principles of right recognize that each person is an agent who can properly claim primary responsibility or authority to define or recognize and revise who she is. This aspect of the self does not describe all that a person is. It does, however, provide the basis for the unity of a self that changes over time.

Sandel provides appealing descriptions of how community can and should partially constitute the person.[23] Nothing in Rawls' analysis, however, rules out Sandel's description of this constitutive role. The principles of justice should be seen only as particular constraints limiting the type of community that fully constituted, "thick" selves can legitimately create. Any community may partially constitute its members, but Rawls' critical project is to identify communities that do so illegitimately. In fact, Sandel's description of properly operating, constitutive communities may presuppose the establishment of Rawlsian justice. Sandel refers constantly to communities that are constitutive of our "*self*-understandings" and he emphasizes that the subject must be "open . . . [to] transformation in the light of revised *self*-understandings."[24] Rawls embodies in his original position, as well as in any resulting principles of justice, the notion of a free and equal person, thus stressing the importance of respecting the individual's autonomy and her right to choose or affirm and to change the particular elements of her self. This emphasis amounts to a claim that the constitutive role of community should operate through individual recognition or acceptance, through *self*-understandings, rather than through imposed understandings. Adoption of Rawls' approach would limit (although necessarily would not eliminate) the degree to which the constitutive aspects of community can be imposed on people. To the degree possible, Rawls' approach requires that community be voluntary.

B. *Sandel's Notion of Group Rights*

Despite his concern with "self-understandings", I read Sandel's argument implicitly to accept a notion that the group or community could justifiably prevail over individual choice or self-definition. This notion, in turn, suggests a conception of group rights. Although Sandel never explicitly refers to group rights, they seem implicit in his talk of

---

[22] For an expanded discussion of this argument, see Baker, *supra* note 5, at 949-72.

[23] *See, e.g.*, M. SANDEL, *supra* note 1, at 172-73.

[24] *Id.* at 172 (emphasis added). *See also id.* at 64, 150, 173, 174.

attributing responsibility to or affirming obligations to groups—to the "family or community or class or nation."[25] As will be noted in Part II, Sandel does not reject the difference principle so much as argue that the difference principle requires the notion of a group subject of possession that can rightfully use one person's attributes for other persons' benefit. This group subject corresponds to a community in the constitutive sense—which is the type of community Sandel consistently suggests exists. I will argue in Part II, however, that this is the wrong way to view Rawls' argument for the difference principle.

Sandel constantly emphasizes Rawls' commitment to the priority or primacy of the self.[26] Sandel then suggests that the self is, in fact, largely constituted by the group—by history and culture. The key difference between Rawls and Sandel, however, is not over whether the individual will find her identity partially constituted by her values, character, and community. Rawls' claim of priority of the self is not a logical, historical or empirical claim—in these senses clearly the group is prior to our concept of the individual. Rather the priority of a choosing self that Rawls asserts and Sandel denies amounts to Rawls' claim that the collective should treat the individual as having moral responsibility with respect to her constitutive attributes, and as having the freedom to affirm or resist possible identities. Rawls asserts that the self should be "prior" for purposes of moral argument. The issue is whether the person has a right to expect that the constitutive force of character, community, and history involve her *self*-understanding or noncoerced understanding; or whether, on the contrary, the constitutive status of these collective aspects of the self means that the group or community can properly and legitimately override individual understanding, individual reflection, and individual rights.

Social thought and societal institutions historically have most commonly embodied the second interpretation—Sandel's interpretation—of the constitutive role and ethical priority of community. Therefore, the liberal's assertion of the moral priority of the individual must be a claim that this priority is the product, not the foundation or even a constant, of history. Despite the frequent failure of liberalism—that is, individual rights theorists—to understand the role of community and its related misguided reification of private property, liberalism's emphasis on the individual was, historically, a revolutionary advance. The theoretical and institutional recognition of the moral priority of the individual is the result of historical moral development. This moral priority,

---

[25] *Id.* at 62-63.
[26] *See, e.g., id.* at 19.

however, properly does not operate to determine the total content of the social world or to eliminate the constitutive communities that Sandel describes as important to the "thick" or actual self. Rather, the priority operates as a restraint on the form these communities can properly take. These communities must respect the autonomy of the person and the person's right to take responsibility for herself. (Note that having this type of right does not imply when, or if, or how, it should be exercised—those questions raise further, more complicated issues.)

By incorrectly implying that Rawls rules out any constitutive role for communities, Sandel was able to carry out his critique without facing the question of how this constitutive role of communities should operate. For example, Sandel does not explicitly claim that all acceptable theories would be unable to specify any general principles that would limit the manner in which the community can properly or legitimately constitute itself. Thus, Sandel does not explicitly defend placing the group above the individual or group rights above individual rights. Nevertheless, these priorities harmonize with his analysis. In contrast, these priorities would be alien to a Rawlsian analysis, which would necessarily derive any group rights from individual rights.

II. The Person and the Difference Principle

Sandel uses Rawls' own statements to argue that Rawls requires a group subject in order to justify the difference principle. Sandel emphasizes that Rawls does, and must, view people's talents not as their own, but as common assets.[27] The difference principle does not allow a person to keep all that flows from the exercise of her talents. If Rawls acknowledged that a person's talents and everything that results from the exercise of her talents were her own, then, in requiring social institutions that distribute benefits according to the difference principle, Rawls would be treating people as means, not ends.[28]

In order for a person's talents to be common assets, Rawls' person must be very sparse. Abstracted of all personal qualities, the Rawlsian person certainly does not correspond to how we view ourselves. But Sandel argues further that stripping the person of all talents is not enough to justify the difference principle. Rawls needs a further argument. Even if a person is not in any way constituted by her innate qualities, it does not follow that these arbitrarily distributed traits should not be left where they fall. Viewing a person as independent of her "natural" qualities or talents implies nothing about whether she

[27] *See id.* at 66, 70-78, 101-03.
[28] *See id.* at 78-80.

should prevail in her claim to possess those talents and all the benefits that flow from her exercise of them. Rawls' abstract person does not lead to any particular distributive principles. In order to successfully complete the argument for his difference principle, Rawls must postulate a group or community subject to own or possess these common assets.[29] Since Rawls' original position, and his entire method, depend on the priority of the denuded and individual moral subject, Rawls ultimately cannot justify the difference principle within his own framework.[30] Justification requires the priority of a group or collective subject.[31]

Here Sandel borrows heavily from Robert Nozick's *Anarchy, State, and Utopia.*[32] Nozick argues that if the talents or natural assets with which people are endowed are arbitrarily distributed, people may only "have," not "deserve" their natural assets. Still, this "having" can be enough for people to be entitled to their natural assets and to whatever flows from them. Since holdings flow from natural assets (presumably Nozick means "from the exercise of these assets"), people are entitled to their holdings.[33] Thus, Nozick concludes,the difference principle is unjust. Or, as Sandel says, justification of the difference principle requires acceptance of a group subject "entitled" to the natural assets and a correspondingly denuded conception of individual selves. In these claims, Sandel is effectively "[p]laying Nozick off against Rawls."[34]

Sandel's play does not work, however, because Nozick's argument itself fails. The claim on which Nozick's argument depends is that material holdings and, therefore, entitlements can flow solely from people's exercise of their natural assets. This claim is certainly wrong,[35] and as it falls, so does Sandel's critique.

Although some physical transformations of the world may "flow" directly from a person's activity, the only things that the actor "naturally" possesses that flow directly from the exercise of her talents are the experience and memory of the activity itself—neither of which Rawls even considers transferring from the actor to someone else.[36] All

---

[29] *See id.* at 96-97, 101-03.

[30] *See id.* at 80.

[31] *See id.* at 101-03, 148-49, 152, 161, 178.

[32] R. NOZICK, ANARCHY, STATE AND UTOPIA (1974).

[33] *Id.* at 225-26 (acceptable argument G).

[34] M. SANDEL, *supra* note 1, at 77.

[35] *See* Baker, *Counting Preferences in Collective Choice Situations*, 25 UCLA L. REV. 381, 398-99, 402 n.63.

[36] Of course, even this "naturalness" follows from a particular conception of the person that we could imagine to be different but that in this respect both I and, I expect, you accept.

other holdings or entitlements that Nozick describes as flowing directly from a person's natural assets will necessarily depend upon the interaction of the individual's activity and the institutional structure (or system of accepted understandings) within which that activity occurs. The determination of who possesses particular attributes is irrelevant to the question of whether and how the exercise of those attributes should be rewarded, or to the question of what entitlements or valid claims flow from their exercise.

Moreover, possession of particular talents or attributes is irrelevant to the choice of criteria for evaluating institutional structures. A person's claim to her bodily features or talents does not give her a basis on which to make persuasive claims concerning the appropriate institutional structure in which these talents are exercised and from which possessory claims arise. Rawls' theory and his difference principle are concerned with the structure in which people act, not with the distribution of natural assets. Thus, nothing in the argument for the difference principle prevents Rawls from viewing a person's talents (as well as her ends, commitments, community, and history) as partially constitutive of the self, certainly of the "thick" self of real life. Whether or not such attributes are partially constitutive of the self is simply irrelevant for the issue Rawls is examining.

If Sandel were right that Rawls takes specific, individual attributes to be collective assets, and that Rawls is committed to a group subject entitled to the rewards of their exercise, then we should expect Rawls to call for more than relatively egalitarian rules in redistributing material or social benefits. Collective concerns should also determine the use of these attributes. For example, Rawls should require those people who hold scarce or valued talents to exercise them for the benefit of all. Loafing would not be an option. A collective subject clearly could conscript effort from some of its parts to benefit other parts. Rawls, however, never calls for this mandatory practice. In fact, such a mandate, which Sandel's interpretation would allow if not require, would presumably violate Rawls' first principle of maximum equal liberty.

Sandel makes much of Rawls"s statement that "[t]he difference principle represents, *in effect*, an agreement *to regard* the distribution of natural talents as a common asset . . . ."[37] "To regard," however, implies viewing the distribution of talents in a particular way for some purpose. The purpose, and the only purpose, for which Rawls "re-

[37] M. SANDEL, *supra* note 1, at 77 (quoting J. RAWLS, *supra* note 2, at 101) (emphasis added).

gards" natural talents as common assets is for choosing or justifying criteria for evaluating social institutions—for choosing the principles of justice. Regarding talents as collective assets for this purpose alone is merely a way of saying that the possession of a talent does not provide an individual with any basis for a claim that she should be favored by the choice of institutions. Rawls' statement merely emphasizes that the possession of particular attributes does not justify a claim that the institutional framework, in which talents are exercised, ought to take a particular form. No claims about the proper institutional form and, thus, no claims to specific holdings, flow naturally from a person's possession or exercise of specific talents. In contrast, an assertion that an individual's claims to property arise merely from engaging in particular activities would not only make no sense but, if the assertion could be given content, would imply that those who are best able to engage in these activities have priority over others. In this respect, Nozick's approach accepts domination by treating those who only have disfavored talents as means to the ends of those with the favored talents.

Institutional design, of course, inevitably favors people with some personal attributes rather than others. The particular form of favoritism, however, is not an automatic result of the biological or social distribution of personal qualities. Rather, favoritism occurs only because of the establishment of a particular set of institutions. Rawls argues that in choosing between alternative institutional arrangements, or in evaluating institutions, people have a claim to be respected as equals. Rawls justifies favoring people with particular attributes not as a goal, or as right in itself, but as a means to achieve just and collectively desirable results. *To regard* the talents, *in effect,* as common assets, but only for purposes of institutional design, means that no one will be arbitrarily favored or subordinated in the inevitably collective and historical process of deciding what claims to recognize and what activities to favor. In other words, contrary to Sandel's assertion, the argument for the difference principle does not require or imply a collective subject.

Since rights flow from a person's activity only within some collective structure, the crucial issue becomes what collective structure do we find acceptable, what claims should we honor? This raises the question of the relevance of various aspects of the person for justifying the social structure. Which aspects of the person are relevant to the justification and what role do they play? Do different aspects of the person relate to different aspects of the justification or relate to it in different ways?

One plausible answer is that all aspects of the person are potentially relevant to all questions concerning the proper form of the social structure. A social structure within which we live and which we consti-

tute should reflect all that we are. If this answer were correct, Rawls would require a full theory of the person. But this answer is wrong. We often conclude that various aspects of what we commonly consider to be a person do not properly contribute to the justification of social policy. For example, we commonly consider a person's preferences and personality type to be aspects of the person. Nevertheless, we may conclude that people's racist preferences, or their personal "need" to dominate, do not play any proper role in justifying social structures, even if these preferences may help explain oppressive aspects of existing structures.[38] Thus, a more plausible—and I think better—answer is that different aspects of the person, or of the human condition, may relate to different questions about the social structure. Aspects that are relevant to general principles that constrain the set of acceptable structures of interaction may be a smaller set than those aspects that are relevant to the choice or evaluation of specific institutions, or for action and politics within that structure. Since Rawls is concerned only with basic principles of justice, he need only identify those aspects of the person, or of the human condition, that are relevant to the content of these basic principles.[39]

If not all aspects of the person should go into justifying the institutional structure, what aspects should? Or, more specifically, what aspects should be relevant for those issues that Rawls claims that the principles of justice settle?[40]

Two observations may help answer this question. First, in order to argue for general, abstract, universal principles of justice,[41] Rawls must

---

[38] *See* Baker, *supra* note 35; Ely, *Professor Dworkin's External/Personal Preference Distinction*, 1983 DUKE L.J. 959.

[39] Sandel would claim, at least, that the original-position lens shows all the aspects of the person that a Rawlsian can consider morally relevant. He repeatedly quotes Rawls' statement in a 1975 article: "That we have one conception of the good rather than another is not relevant from a moral standpoint." *See, e.g.*, M. SANDEL, *supra* note 1, at 74, 165, 180 (quoting Rawls, *Fairness to Goodness*, 84 PHIL. REV. 536, 537 (1975)).

As I read Rawls, this statement does not imply that a person's conception of the good is not an important matter that can be subject to praise or criticism even apart from the conception's congruence with principles of justice or universal moral principles. On the page following the one Sandel quotes, Rawls makes clear that his claim is that a person's conception of the good, which will necessarily be affected by arbitrary contingencies (as well as by the person's conception of the right), is irrelevant from the standpoint of the original position. *See* Rawls, *supra*, at 538. Whether the original position exhausts the moral standpoint or only provides the basis for considering certain specific foundational issues is not clearly taken up by Rawls. Although his Kantianism might counsel the other way, I would argue that nothing in his general methodology rules out the second possibility.

[40] Elsewhere, I argue that the number of issues settled may be fewer than some readings of Rawls would suggest. *See* Baker, *supra* note 5, at 947 n.43.

[41] Because of my own tendency to conclude that *we* should be committed to the

argue that certain features of human beings, or of the human condition, are universal, either in fact or aspiration. These universal features could be either qualities common to us all—such as our capacity for choice, reflection, and responsibility—or features that we must attribute to anyone whom we can justifiably expect to accept the rules of the structure, or to anyone with whom we could engage conversationally in justifying our claims.[42] Whichever interpretation one accepts, my point is that principles for evaluating or designing human society can only be general or universal in light of something general or universal about the "facts" or practices on which they are based. Second, Rawls must argue that the universality of these features justifies certain constraints or requirements in the design of our institutions. Only those universal features that are presupposed in people's making claims on one another or, more generally, are presupposed in moral action, seem relevant. In this sense, Rawls' theory again points more toward a theory of intersubjectivity than of individual consciousness.

The presence of such features and the development of their implications are key to Rawls' theory. Thus, the argument for general principles of justice can be broken down into more specific questions. Can we properly postulate such common or universal features either of the person or of morally acceptable contexts of interaction? If so, do these common features merge with other more particular features in any justification of social structures such that no specific conclusions follow from the common or universal features? Or, alternatively, do these features have some preeminent role that can justify giving priority to those principles that best respect or embody them? Both the argument for general principles of justice, and the argument for the priority of justice, depend on the answer to these questions. On my reading, Sandel did not face these crucial issues, issues to which I will return in Part V.

III. Distributive and Retributive Justice

In his critique of Rawls' analysis of retribution, or punishment, Sandel repeats his mistake of failing to distinguish between the aspects of a person relevant to choosing or evaluating frameworks of interac-

---

universal validity of the principles of justice that I think can be developed using Rawls' methodology, I may overread Rawls on this point. *See, e.g.*, Rawls, *supra* note 11, at 518 ("[W]e are not trying to find a conception of justice suitable for all societies regardless of their particular social or historical circumstances. . . . How far the conclusions we reach are of interest in a wider context [than that of a modern democratic society] is a separate question").

[42] *See* J. Habermas, Communicative Action, *supra* note 14. *See also* B. Ackerman, Social Justice in the Liberal State (1980).

tion, and those relevant once the framework is chosen.

Sandel first notes that Rawls rejects the desirability and coherence of basing distribution on "moral desert"—a term that Sandel treats as referring to qualities antecedent to public institutions.[43] This follows for Rawls because moral worth, on which desert could be based, is determined by conduct within the structure or context in which a person finds herself. Thus, neither moral worth nor any other feature can provide a preinstitutional basis for determining desert.[44] Sandel and Rawls are both, however, inclined to describe claims arising from actual institutions as expectations, not deserts. Rawls does not even bother to defend any particular definition of moral worth. His one suggestion, however, is that it be viewed as an effective desire to comply with the rules of a just society.[45] Given this conception, moral worth is subsidiary or subsequent to the principles of justice. It can have no bearing on distributive principles. (It would be an odd principle that says: distribute more to those with moral worth, that is, those who have an effective desire to comply with principles that distribute more to them.) More important to Rawls, people in the original position, not knowing if they themselves would be virtuous, would not choose rules that aim at rewarding virtue. Rawls argues that, instead, they would be motivated to choose something like the difference principle. As a consequence, people's holdings or entitlements should reflect legitimate expectations based on postinstitutional facts. These expectations would reflect a person's circumstances and the consequences of her activities within a just institutional structure.

Despite Rawls' rejection of preinstitutional merit for purposes of distribution, Sandel claims that Rawls has a preinstitutional notion of desert for purposes of retributive justice.[46] Sandel concludes that this

---

[43] *See* M. SANDEL, *supra* note 1, at 85-88.

[44] Preinstitutionally, even the relative market value or social significance of people's natural attributes is entirely indeterminant. In this sense, people are roughly equal before any institutional structure that exists, as they are in Rawls' original position. (Note that no claim is made that there could be such preinstitutional people. The imagery is intended to dramatize how much is subject to humanly created and changeable structures.) For the reasons Rawls develops, this preinstitutional equality would not imply the necessity of an equal postinstitutional distribution. Moreover, in a Rawlsian society, as in our own, the supply and demand for various talents will affect the actual distributions if wealth and opportunities, but surely one would not suppose that changes in supply and demand correspond to changes in a person's moral worth. *See* J. RAWLS, *supra* note 2, at 311.

[45] *See id* at 312.

[46] Sandel claims that Rawls is committed to the view that no one can be said to deserve anything. *See* M. SANDEL, *supra* note 1, at 85, 87, 88. This claim, of course, follows only given Sandel's reservation of the notion of desert and moral worth to preinstitutional facts rather than to legitimate expectations based on social institutions. (Sandel, however, quotes Rawls' statement in which Rawls refers to "legitimate expec-

difference between the assumptions on which Rawls bases distributive and retributive justice show a fundamental confusion in Rawls' theory.[47]

Rawls, however, does not have a preinstitutional theory of retribution, except in a very limited and, for Sandel's purposes, irrelevant way. Rawls assumes that a common feature of people or, more specifically, a feature that our political and moral interactions must assume, is a person's right to be treated as an autonomous person with the capacity to act responsibly. Given this attribution of responsibility and autonomy to the abstract individual, a theorist can abstractly and preinstitutionally conclude that a person's violation of the rules of just institutions is presumptively a mark of bad character.[48] In contrast, given that particular talents are arbitrarily distributed, the theorist cannot abstractly conclude that for a person to act in a way that just institutions reward provides a basis for anything more than legitimate expectations. In this sense, there is a moral basis for retributive justice that is lacking for distributive justice. Still, both permissible retribution and proper distribution depend on institutional structures.[49] Neither

---

tations established by social institutions" as "this sense of desert." *Id.* at 72 (quoting J. RAWLS, *supra* note 2, at 103)). Sandel also attributes to Rawls the claim that retributive justice refers to desert and intrinsic moral worth. On this basis, Sandel suggests that Rawls has contradicted himself. *See* M. SANDEL, *supra* note 1, at 90.

[47] *See* M. SANDEL, *supra* note 1, at 90-92.

[48] Sandel quotes Rawls: "Thus a propensity to commit such acts is a mark of bad character." *Id.* at 90 (quoting J. RAWLS, *supra* note 2, at 314-15). Several caveats need emphasis in order to avoid misinterpreting the argument in the text as applied to the criminal justice system. First, the argument applies to people within a basically just order. As applied to the existing order, its implications are much more ambiguous and no obvious conclusions can be drawn. Second, even in a just society where the criminal behavior would presumably be a mark of bad character, Rawls' argument does not imply that there are two sets of people, a few who are marked with a bad character and the rest of us who are not. Given that we are all imperfect beings, we all will, on occasion, be subject to both criticism and praise. Third, the argument about individual responsibility and autonomy does not speak to how we should respond to the person subject to criticism—whether our response should exhibit compassion, mercy, or an effort to give aid. Certainly, my second point implies the evil of a harsh, self-righteous response. Fourth, the justification for the way the Rawlsian argument ignores the determination/free will debate lies in the need to show respect for ourselves as moral beings—as free and equal persons—and hence the desirability of designing social institutions in a form that is consistent with the grammar of our practices as moral beings. *See, e.g.*, Strawson, *Freedom and Resentment*, in STUDIES IN THE PHILOSOPHY OF THOUGHT AND ACTION 71-96 (P.F. Strawson ed. 1968). To the extent that from the perspective of sociology we diverge from these free will premises, our design of social institutions can properly take this sociological information into account. This design, however, must do so in a way that is consistent with our practices that involve the perspective of freedom and morality, that is, that assume agency and free choice.

[49] Rawls does say that the criminal law upholds "basic natural duties." J. RAWLS, *supra* note 2, at 314. The key point, however, is that the criminal law sets standards about impermissible behavior that all people can properly be expected to

can be identified preinstitutionally. Robbery, for example, cannot be identified absent social practices that recognize property.

Sandel emphasizes Rawls' statement: "*To think of distributive and retributive justice as converses of one another is completely misleading and suggests a moral basis of distributive shares where none exists.*"[50] Contrary to Sandel's interpretation, Rawls' statement does not imply that "punishment [will tend to] correspond to [preinstitutional moral notions.]"[51] Rather, the statement reflects two fundamental differences between retributive justice and distributive justice.

First, the key issue of distributive justice is to determine what entitlement or property rules and distributive practices we should have. In contrast, the key issue of retributive justice is how we should respond to violations of the rules that define just institutions. Thus, the two are not converses of each other, but do entirely different work. Distributive justice is, in this sense, more basic in that it poses the question of the proper content of those social institutions to which retributive justice has relevance.

Second, and here the different moral bases of the two come into focus, distributive and retributive justice differ in the moral force of their directives. Distributive justice determines what we can legitimately expect from actions that are morally permissible. The actions are permissible in that, at least without more particularized knowledge of the person and her circumstances, we cannot morally condemn or praise a person for being able to, and then taking, the particular action that gives rise to the legitimate expectation. In contrast, retributive justice specifies actions that, given just social institutions, we can morally condemn. This condemnation does not imply that the occasions for condemnation and retribution can be determined preinstitutionally—rather the occasions are often dependent on first establishing a just distributive order. The condemnation does reflect, however, the preinstitutional judgment that people should be criticized for violating the rules of a just order.

The difference between the moral content of this retributive judgment and the nonmoral basis of the distributive order's legitimate ex-

---

meet. (Although, after setting such standards, an account of people's real world conditions may properly influence punishment decisions. *See id.*) In contrast, distribution refers to other types of facts. The assumed general capacity to conform to criminal law provides no reason to think everyone has the same capacity to engage in economically rewarding behavior which normally requires talents (which are not evenly distributed) as well as the capacity of choice that is implicit in this notion of autonomy.

[50] M. SANDEL, *supra* note 1, at 90 (quoting J. RAWLS, *supra* note 2, at 314-15) (emphasis added by Sandel).

[51] M. SANDEL, *supra* note 1, at 90.

pectations specifically reflects the original position methodology. The choice of the proper distributive order does not involve any evaluation of the natural worth of particular personal attributes, but is a response to each person's claimed right to be permitted to try to lead a fulfilling life. Thus, the distributive order's response to the exercise of natural attributes implies nothing about the moral worth of the actor. If the social order does not reward a particular activity highly, it would not mean that the person was less morally worthy. The contribution of an activity to the development of a better social order is often not proportional to the amount the activity is rewarded. The lack of a large reward should only mean that a larger reward is not justified by its contribution to a good and just society—for example, the reward is not justified by the consequences of creating different behavioral incentives.

In contrast, retributive sanctions do imply moral evaluations. The original position is premised partly on the moral importance of treating the person as a responsible agent of choice. From the perspective of the original position, this assumed capacity for choice provides the moral basis to condemn the person who violates the rules of a just order. This condemnation does not flow intrinsically from any specific behavior or particular personal attribute. It follows because of a quality attributed to everyone—the capacity for choice within a just structure. Rawls' theory need not even assume that this capacity is empirically operative. The capacity is only assumed as a necessary premise for many of our fundamental moral practices and aspirations—for the idea of legal obligation or, alternatively, for noncoerced conversational agreement. Thus, the assertion of a moral or preinstitutional difference between distributive and retributive justice does not show a confusion in Rawls' theory of justice or his conception of the person. Rather, the original position method both requires and explains the difference.

Sandel's confusion concerning Rawls' distinction between distributive and retributive justice, like his tendency to see Rawls as being committed to a notion of the self that is abstracted from particular personal qualities, illustrates a more fundamental misunderstanding of the Rawlsian enterprise. Rather than looking back through the lens of the original position to see a picture of the basic, constitutive elements of the Rawlsian theory of the person, I have argued that one should see only those elements of our notion of our selves, or those elements of human interaction, that justify various constraints on the justifiable form of our collective activities. A successful critique of Rawls' methodology must show either that no general elements of selves or of human interaction should be assumed, or that these general elements do not justify general constraints.

IV. JUSTICE AS A VICE

Sandel's assumption of the priority of the group, and possibly of group rights, is evident in his criticism of the primacy or priority[52] that Rawls accords to justice, criticisms that he caps with the claim that justice may sometimes be a vice.[53] Sandel's argument, however, relies on mistaken claims concerning the empirical circumstances that Rawls must assume[54] in order to argue for the primacy of his two principles of justice. Sandel also relies on a mistaken notion of the social practices that the two principles of justice require.

Rawls recognizes that the priority of justice depends on the existence of certain empirical conditions—for example, moderate scarcity and conflicting desires for the use of resources. Sandel, however, interprets Rawls to require that the empirical circumstances make justice burdensome. The priority of justice will exist, Sandel argues, only where a high degree of scarcity and conflicting demands, together with an absence of benevolence, make justice "the most pressing social priority."[55]

Sandel's argument confuses two senses in which justice may be primary. Its primacy could mean that establishing justice is the most difficult and pressing—most burdensome but most important—objective for a particular group. Sandel seems to give "priority" this meaning.[56] In contrast, for Rawls, the primacy of justice relates to ordering. Justice should prevail if it conflicts with other institutional concerns. These conflicts, however, may not seem severe; the conflicts need not even exist. Clearly, Rawls hopes and gives reasons to expect that, in a just society, the demands of justice will not weigh heavily on people. Rawls argues that, particularly in a well ordered society, the "right" will be generally congruent with people's conceptions of the "good." Under these circumstances, although justice will be basic or fundamental, it will not be felt as pressing. Sandel's failure to see that Rawls' claim

---

[52] Primacy might refer to "weight"—that is, relative importance—while priority might refer to "ordering"—that is, determining which prevails in case of a conflict. Priority might be more suggestive of a temporal or logical order. Sandel, however, does not appear to distinguish the two terms consistently. *See, e.g., id.* at 2, 7, 9, 14, 15, 17, 21, 30 (primacy of justice); 7 (primacy of moral subject); 10, 16 (priority of justice); 8-9, 15, 20, 21 (priority of moral subject).

[53] *Id.* at 34.

[54] Sandel notes that Rawls' argument can escape dependence on empirical circumstances, but only by developing other problems, *see id.* at 40-46, in particular, by adopting an impoverished and unappealing picture of the moral subject. *See id.* at 47-65. Sandel is mistaken to suggest that Rawls either does or needs to treat the circumstances of justice as existing only in the original position. *See id.* at 41.

[55] *Id.* at 30.

[56] *See id.* at 30-36, 173.

that justice is a first virtue is an ordering claim, not a claim about its burdensomeness, is hard to explain. This mistake would, however, justify Sandel's wish to focus attention on the empirical realm in which people presently exist and in which the dominant constitutive elements are history, culture, and various types of groups.

Sandel continues his analysis by arguing that where a spirit of fraternity or benevolence prevails, justice will not and possibly should not be engaged. He argues that, in such a situation, a move toward increasing the justness of the group may actually worsen matters. Therefore, in these circumstances, justice should count as a vice.[57] This claim, however, presupposes a mistaken view of the Rawlsian notion of justice.

Justice is a virtue of institutions, or of frameworks of interaction.[58] It concerns how the general structure of basic institutions distributes advantages and burdens. Just institutions do not prohibit people from acting out of various individual motives that are consistent with a sense of justice, although not identical to it. The dominance of a spirit of fraternity and benevolence as individual motives for action within a community is perfectly consistent with the primacy of justice. The presence of these sentiments might lead the prosperous to feel generous toward the claims of the worst off and make the fulfillment of their claims more likely. Thus, the presence of benevolent motives may even make the justice of basic institutions more likely.

Moreover, the requirement that institutions be just does not require that everyone actually get precisely her due, nothing more and nothing less. Justice does not require the due process version of hell. The justness of institutions says nothing about whether a person will, or should, push to the limit her rightful claims to various advantages. Sandel reads Rawls' exhortation to act "on the principles of right and justice as having first priority"[59] as requiring that we "act out of a sense of justice."[60] But neither Rawls nor the priority of the principles of justice require that we assert all possible claims to get our due or that we force resources on those who would prefer to do without. At most, Rawls' theory of justice requires that we support institutions that *permit* people to obtain what justice would make available. A pervasive spirit of benevolence and fraternity, even if it leads some of the worst off to accept less than they have a right to demand, does not represent a decline in the justness of the social order. As long as the basic institu-

---

[57] *Id.* at 32-35.
[58] *See* J. RAWLS, *supra* note 2, at 3.
[59] M. SANDEL, *supra* note 1, at 23 (quoting J. RAWLS, *supra* note 2, at 574).
[60] M. SANDEL, *supra* note 1, at 22.

tions would allow those worst off successfully to demand more, their willing acceptance of less is perfectly consistent with a just order. Their acceptance of less can be seen, roughly, as an exercise of their liberty to "spend" or share some of their possible claims in a way that reflects their notion of the good, their sense of virtue, or their notion of their own identity. A just order still permits vows of poverty.

Given this rather obvious consistency between justice and a spirit of benevolence or fraternity, why does Sandel conclude that the two conflict? His conclusion would make sense in a particular factual scenario, but only if Sandel adopts a particular set of ethical assumptions.

As noted above, a person's benevolent communal sentiments, religious principles, or other rational life-plans might cause her to choose a lesser position or a smaller set of benefits. This situation is not inconsistent with the priority of justice as long as established institutions would allow the person successfully to claim a more equitable position or set of benefits. No conflict exists between justice and communal sentiments. The structure still recognizes the priority of justice, since the individual's just claim, if made, would prevail.

This just order, however, should be sharply distinguished from the situation where, whether out of a sense of personal identity, benevolence, community, or tradition, a group of people accepts a form of human association that forcibly subordinates or unequally limits the options of some people in order to advance the ends of others, or of the group. In this situation, communal sentiments would require that certain people not be able to obtain as much as they could under just institutions. The group would deny some people's claims to the opportunities that just institutions would provide. This form of human association does violate the requirements of justice. Even if their communally-based sentiments lead the subordinated people to accept this form of interaction, the content of what they accept is the notion that some people have lesser claims. Under these circumstances, anyone's demand for a just order would upset the previously harmonious communal order. A more just order would undermine the prior sense of community. If Sandel equates virtue with communal sentiments, then a conflict between justice and virtue would exist in this situation.

Sandel can identify virtue with communal sentiments and favor this virtue over justice only by making a particular ethical assumption: he must assume that it is proper for the group to subordinate the individual—that group rights may properly prevail over claims that are based on the ethical priority of the equal, autonomous individual. Only under an assumption of the priority of the group could an increase in justice ever be a vice or a decline in the moral status quo.

The proper resolution of a real conflict between justice and specific forms of communal solidarity presents a difficult issue. Sincere and well-meaning Westerners often perceive this conflict in non-Western communities, and they are divided as to whether justice or communal tradition ought to prevail. Rawls, at least on a universalist interpretation, implicitly asserts the primacy of justice.[61] Sandel's choice of examples distorts and, in the end, avoids the issue. Sandel emphasizes the appeal of fraternity in situations in which justice and fraternity do not really conflict. He describes situations in which both are present but in which the requirements of justice are not burdensome and therefore do not seem most important; or he describes situations in which people do not make claims based on justice. Nevertheless, to the extent that Sandel seriously claims that a just Rawlsian social order can be a vice, he must necessarily assume situations in which virtue requires the subordination of some people, an assumption that implies the ethical priority of the group over the individual. For purposes of evaluating the structure of interaction, Sandel must assert the priority of a group right over individual rights. The Rawlsian claim with respect to this central but difficult issue should be that our ethics, with its commitment to individualism is merely the historical product of Western moral development, properly recognizes the necessity of undermining the structures of domination and subordination that Sandel's formulation would find acceptable.

V. THE PLAUSIBILITY OF A TWO-LEVEL THEORY

If we were to accept any principles, such as Rawls' principles of justice, as having some sort of universal validity, presumably this acceptance would derive from the relation of the principles to qualities of humans or of human interaction that we consider to be in some sense universal. The plausibility of the claim that we recognize such universal qualities is critical to Rawls' argument. Nevertheless, neither the claim of universality for such qualities nor the claim of universality for the principles of justice should be read too broadly. An obvious problem with Sandel's philosophical anthropology, which has been discussed, is the suggestion that he uncovers Rawls' complete theory of the person.

---

[61] This claim of primacy obviously could be viewed as "cultural imperialism." I recognize that, although the implications of and alternatives to this position need greater attention. The claim, however, may not justify—in fact, it may rule out—typical, imperialistic methods of implementing imperialistic claims. *See* Baker, *The Process of Change and the Liberty Theory of the First Amendment*, 55 S. CAL. L. REV. 293 (1981). Moreover, a different problem is that the Westerner's perception of conflict is often distorted and inaccurate.

Even the more limited claim that Sandel finds the core, although not the entirety, of Rawls' conception of the person is not persuasive. Clearly many of the most significant, defining, and meaning-giving qualities or aspects of people's lives are not recognized to be universally present. People's personal or cultural particularity or uniqueness is normally crucial to their sense of identity, importance and meaning.

Just as the attribution of universal qualities to humans, or to human interaction, does not imply a complete theory of the person, these qualities also do not imply that the corresponding principles of justice will totally prescribe the proper social structure or individual behavior. Many aspects of both the social structure and individual behavior do and should respond to variable, nonuniversal qualities. If the principles of justice operate only as a limit on permissible structures of interaction, there is no basis to fear denying or completely suppressing the variable features of history, culture, or people's lives.[62]

A variety of other reasons reinforce the conclusion that general principles could not completely determine the proper structure of society. The nature of knowledge or understanding, more specifically, the inevitably open texture of concepts or values, implies that the necessarily open textured-principles could never be completely determinative even with respect to the basic structure of interaction. The point about open texture merely underlines our recognition that the meaning of the fundamental values or principles, their relation to the presumptively universal features of our situation, and these universal features themselves, will all be human constructs that capture some aspects of our history, experience, and aspirations rather than mirror any essential reality. More relevant to the argument here is that the conception of universal principles claims derivation from the notion of free and self-determinative agents of choice, whose "thick" selves will necessarily be imbedded in communities. This conception would be inconsistent with its own premise of a free agent of choice if it claims that the principles completely determined the basic structure. The principles are better conceived as vetoing those arrangements or structures that would undermine or deny the universal attributes. Respect for free and autonomous agents requires that the nature of just groups and social institutions ought to be, as existing groups inevitably are, variable. Room must be left for politics.[63]

Thus, if the universal principles have any sort of priority, it could

---

[62] Another way of describing this notion of the principles operating only as vetoes, suggested to me by Eric Hoffman, is that the principles amount to a theory of injustice, not to a theory of justice.

[63] *See* Baker, *supra* note 5, at 949-63.

only be to restrict the range of permitted choices concerning the structure of interaction. Although a defense of these universal principles would claim that they reflect general features that we, in some sense, recognize in the human situation, the defense need not claim that these general features are necessarily most important for our self-definition. The defense must claim only that these general attributes are the most relevant or most important for certain purposes. Specifically, it must claim, first, that these general features exist, either empirically or as attributes that we recognize as presupposed in basic human practices or as aspirations implicit in those practices. Second, the defense of these universal principles must claim that the role of these features in people's lives or interactions relates them to the content of foundational, universal principles. Third, the defense must claim that these general features have a moral importance such that we should not permit concerns based on more particular features to undermine the corresponding general principles—that is, the principles of justice should have a particular, ordering priority. This theory, however, also recognizes the existence of a second level of more particular qualities, whether based on choice, cognition, self-reflection, or history and community. These more particular and concrete qualities are necessary to give more complete content to the social world, to our understanding of ourselves, and to our activities.

This type of two-level theory, which is crucial for the Rawlsian approach to justice, is certainly problematic. Sandel's critique of Rawls, however, never explicitly challenges this theory. Sandel could have tried to critique this two-level theory in at least three ways. First, Sandel could have argued that there are no general features of humans or of human interaction on which to ground universal principles. Second, Sandel could have argued that one cannot persuasively move from the role that these universal features play in our lives or interactions to any general or universal regulative principles—which is the step that Rawls implicitly uses his original position methodology to take. Third, Sandel could have argued that a political or ethical theory is misguided to claim that some aspects of the person or of human interaction have priority even for certain issues. The second and third lines of argument could even admit the existence of universal features of the human situation, like people's inevitable striving for and need for both individuality (including privacy and autonomy) and community (including recognition by others). This could be admitted because these arguments could still claim that the appropriate scope or force of these universal features of human existence is always very contextually dependent and that other, more particular qualities are always crucially important. Because

of this dependence on context, a theorist or a society could never arrive at anything more than universal principles or values that a person ought to always take into account—or should take into account if the person has adequate time and energy. These values, however, are not ordered in a certain way. As a result, intuitive ad hoc responses and integrations are all that would be possible. This is the intuitionist ethical position. Recognition that this is our situation may be both honest and liberating.[64]

The above arguments, it seems to me, are the type that Sandel must make in order to succeed in his critique of Rawls. Although I am not convinced by these arguments, I cannot adequately respond to them here. I will, however, make three further comments.

First, if the intuitionist ethical position described above is that no one theory will be able to move persuasively from general features of the human situation to general principles that have a particular social or political status—that is, have priority—the most effective response is to do what the intuitionist says cannot be done. Thus Rawls adopts the right strategy. He attempts to find more precise implications of these features of our being that are most deeply rooted in our practices and aspirations and to give more precise content to the resulting seemingly universal principles. To the extent that the intuitionist is unable to effectively criticize such arguments, her position is weakened. (Of course, people will disagree about whether criticisms have been effective.)

Second, the sense that principles of justice, or any universal principles, ignore our particularity, our culture, our community, and our ethic of care, clearly weakens the appeal of such principles. This criticism, however, is blunted by the recognition that the claims put forth on behalf of justice are limited. Properly formulated principles of justice are consistent with and, in fact, contribute to the possibility of a free development of particularity, culture, and community. They leave open most issues of both structure and individual decision.

Third, even if current attempts to give content to universal principles have failed, or even if a convincing theoretical argument could be made that no attempt at enunciating a theory of universal principles could be completely successful, one still might conclude that these attempts to find and give content to universal principles are a useful, intellectually honest, and intelligent aspect of a strategy to achieve moral and political progress. These attempts may be the best means to

---

[64] For defenses of an intuitionist approach from two different political perspectives, see, *e.g.*, Kennedy, *Form and Substance in Private Law Adjudication*, 89 HARV. L. REV. 1685 (1976); Shiffrin, *Liberalism, Radicalism and Legal Scholarship*, 30 UCLA L. REV. 1103 (1983).

engage our always historically bounded reason in the critique and revision of the established order. In criticizing the established order, the attempts to devise and think through the implications of principles that are treated as universals may provide effective leverage against the inherent conservatism of intuitions that are informed primarily by the status quo. At the same time, this use of principles may help us to avoid the arbitrariness and dangers of utopian but unreasoned speculation.

Rawls assumes that social organizations, especially obligatory social orders, ought to respect people's autonomy, the individuals' equal right to or responsibility for choice or discovery. In part, this is apparently because he also assumes that no one can be expected to accept an inherently lesser claim to have her autonomy affirmed than anyone else. Presumably Rawls would argue that these assumptions are basic to our notions of moral interaction or, as Habermas might conclude, to undistorted communicative action. In other words, these premises apply universally to human interaction and, arguably, are universally latent in people's attempt to interact humanly. For these reasons, Rawls presumably would argue that despite whatever particular, historical considerations go into the formation (and continual transformations) of a structure of interaction, the structure ought also to embody these general assumptions. For this reason, Rawls can conclude that these assumptions can be premises from which he can try to formulate universal principles. Or, on a more modest account, Rawls can argue that anyone who accepts these general premises, which will include most people in the post-Enlightenment world, should be committed to any principles that can be derived from them.

Sandel's critique of Rawls closely resembles Roberto Unger's critique of the morality of reason.[65] Unger first shows that the morality of reason responds to only two of four fundamental elements that he identifies in our understanding of the person; those two are the notion that the person has a continuing identity over time and the notion that the self shares a common humanity with others. Unger argues, however, that the abstract timelessness of this morality of reason is destructive of the second two elements, the notions that a person is able to change her ends over time and that the person is a unique, historical individual. The morality of desire, in contrast, responds to the second two elements but is destructive of the first two. Unger then argues that liberal theory

---

[65] *See* R. UNGER, KNOWLEDGE AND POLITICS 49-59 (1975). As used by Unger, the morality of reason, exemplified by Kant, can be seen as involving abstract universals while the morality of desire, often associated with such philosophers as Bentham, focuses on concrete particulars.

is unable to integrate these two moralities—that the abstract humanity implicit in the first two elements is constantly opposed to, rather than integrated with, the concrete humanity implicit in the second two elements. Following Unger's critique of the morality of reason, Sandel can be read to argue that Rawls' conception of the person, at best, only contains the first two elements of an adequate theory of the person.

The most persuasive Rawlsian rebuttal to this critique would be to show that an integration of universal and particular elements is possible, or to show that our interactions can simultaneously involve both, even if our sight only fixes on one at a time. At the collective level, the general or universal constraints placed on an acceptable framework of interaction respond to our constancy over time and our common humanity. For example, attributing responsibility and autonomy to agents implicitly recognizes a constant person who makes choices and who, in this regard, is like other people whose capacity for choice also ought to be respected. The theoretically mandated protection of politics as an essential part of the framework of interaction, as well as recognition of the variability of permissible frameworks, responds to the second two elements—the notions of change over time and of uniqueness.

Presumably a similar integration is conceivable at the level of individual action or individual morality. This integration begins to give content to Rawls' claim, quoted by Sandel,[66] that the principles of right provide for the unity—but, I also emphasize, not the entirety—of the person. The universal principles provide unity both by their priority and by their delineation of the proper realm of the good. They provide unity by recognizing the responsible agent of choice as the center to which the particular chosen or affirmed elements relate. The equally important chosen or affirmed elements provide for the uniqueness and changeability of the self. This rebuttal of Sandel's critique claims that an adequate theory of the person or of society can be, and arguably must be, a two-level theory. An adequate response to some aspects of what we take as central about being a person is a set of general or universal principles that constrains or restricts further choice. An adequate response to other aspects that we also take to be key must involve acceptance of a realm where principles—or rules or choices—are a matter of individual and collective struggle and self-definition in particular historical contexts.

To the extent that these universal and particular elements of the single self are held apart in different realms—the realm of the right or the morality of reason, and the realm of the good or the morality of

[66] M. SANDEL, *supra* note 1, at 21.

desire—a person would not be able to experience the self as a coherent whole. The need for a union of the particular and the universal may be expressed in the wish for personal immortality. More secularly, the need to experience the self as a coherent whole may require somehow having the two elements penetrate or encompass each other. Universality must be made concrete; particularity must avoid being arbitrary. At a formal level, this penetration is partially accomplished by deriving the justification for a particular, communal politics of self-definition from the logic of the morality of reason. From the other direction, penetration is furthered to the extent that a key objective of the morality of desire's particular politics is to find or create greater content for our common existence or to give greater meaning to our notions of equality and humanity. Each concrete, particular individual seeks to contribute to the content of human nature, our universal nature. In practice, one might expect that this integration could occur only through a renewed sense and a new version of community, one that will require greater efforts to maintain openness and dialogue.

Regardless of how this integration occurs, the insight gained from the modern development of an individualistic ethical and political theory is that neither the universalist nor the particular side of the person should be lost, although both are necessarily embedded, in community. In this sense, the liberal split of the self, which Unger rightly criticizes, provides guidance for the nature of community that Unger, Sandel, and Rawls presumably each seek. In other words, attention to these two aspects of the self may provide guidance toward the type of community in which the split is healed or, at least, toward the community that can mediate the conflict and reduce the inherent tension.

Despite sharing this goal of an adequate conception of community, Sandel's notion of community or of a group subject is too unlimited or unstructured. In contrast, Rawls' theory of justice provides the theoretical grounds for demanding, in line with Unger's hopes, that the needed community be open and egalitarian. Both Rawls and Sandel can recognize the crucial importance of community. The contribution of liberalism, as developed by Rawls, is the recognition that an acceptable community must respect the autonomy and equality of the individual. It must conform to principles of justice even as its richness goes beyond them. In other words, the needed community can and must respect both the universal and particular aspects of our conception of the person.

Sandel writes that

> in so far as our constitutive self-understandings comprehend a wider *subject* than the individual alone, whether a family or tribe or city or class or nation or people, to this extent

> they define a community in the constitutive sense. And what marks such a community is . . . a common vocabulary of discourse and a background of implicit practices and understandings within which the opacity of the participants is reduced if never finally dissolved . . . .[67]

This statement about "what marks a community" and the claim that our constitutive self-understandings will comprehend membership in a community both seem right. This statement, however, may be objectionable if "subject" means an entity with the rightful authority to deny freedom or to subordinate some of its own elements in favor of others—if "a wider subject," like "group rights" that are not derivable from individual rights, means an authority independent of the necessity of respecting the autonomy and equality of the individual subject. Then Sandel's invocation of a "wider subject" suggests acceptance of an unjustifiable form of closed or hierarchical community. Given these possible implications of claiming that our self-understandings encompass a "wider subject" rather than merely encompassing community, I find that label doubtful and dangerous. Nevertheless, with that caveat, and contrary to Sandel's assertion,[68] his general notion of community and his claim that our self-understanding is partially constituted by our membership in communities, is consistent with Rawls' analysis and with the type of two-level theory I have described. The consistency between Rawls' abstract principles of justice and our place in variable, historical communities results from a conception of the person and of human interaction that includes several elements. The differing role played by each of the two sets of elements explains the impossibility of arriving at a complete Rawlsian view of the person by looking backwards through the lens of the original position.

Although Rawls can accept much of Sandel's claims about community, Rawls provides an important addition. He tells us something about the appropriate content of these communities. Rawls argues that there are minimal conditions necessary for the community's "common vocabulary of discourse" to be undistorted and for the "background of implicit practices" to be acceptable. He argues that this discourse and these practices must be consistent with the principles of justice. These principles of justice respect—because they follow from—certain features of the common humanity that we presuppose and aspire toward within our interactions.

Sandel's argument that a vision of justice is an incomplete aspira-

---

[67] *Id.* at 172-73 (emphasis added).
[68] *See id.* at 173.

tion should be obvious—but this argument is very different from the further claim that "the vision is flawed."[69] Community may be central to our identity. If so, then the primary implication of the claim of primacy for justice would be, using Unger's language, a claim that an acceptable community must be open and egalitarian, not closed and hierarchical. Sandel's arguments about the nature of the self certainly do not foreclose the possibility that certain universal principles, reflecting fundamental aspects of our interaction, provide guidance about some aspects of the structure of our interaction. Rawls demonstrates the plausibility of this possibility.

[69] *Id.* at 1.